On the Impossibility of Learning the Missing Mass

entropy

MDPI

Article

On the Impossibility of Learning the Missing Mass

Elchanan Mossel [1,†] and Mesrob I. Ohannessian [2,*,†]

[1] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02142, USA; elmos@mit.edu

[2] Toyota Technological Institute at Chicago, Chicago, IL 60637, USA

* Correspondence: mesrob@ttic.edu

† This work was conducted when both authors were visiting the Information Theory Program, 13 January–15 May 2015, at the Simons Institute for the Theory of Computing, University of California, Berkeley.

Received: 18 June 2018; Accepted: 6 December 2018; Published: 2 January 2019

Abstract: This paper shows that one cannot learn the probability of rare events without imposing further structural assumptions. The event of interest is that of obtaining an outcome outside the coverage of an i.i.d. sample from a discrete distribution. The probability of this event is referred to as the "missing mass". The impossibility result can then be stated as: the missing mass is not distribution-free learnable in relative error. The proof is semi-constructive and relies on a coupling argument using a dithered geometric distribution. Via a reduction, this impossibility also extends to both discrete and continuous tail estimation. These results formalize the folklore that in order to predict rare events without restrictive modeling, one necessarily needs distributions with "heavy tails".

Keywords: missing mass; rare events; Good–Turing; light tails; heavy tails; no free lunch

1. Introduction

Given data consisting of n i.i.d. observations $X_1, \cdots, X_n$ from an unknown distribution p over the positive integers $\mathbb{N}_+$, we traditionally compute the *empirical distribution*:

$$\hat{p}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i = x\}.$$

To estimate the probability $p(E) := \sum_{x \in E} p(x)$ of an event $E \subset \mathbb{N}_+$, we could use $\hat{p}_n(E) := \sum_{x \in E} \hat{p}_n(x)$. This works well for abundantly represented events, but not as well for rare events. An unequivocally rare event is the set of symbols that are *missing* in the data,

$$E_n := \{x \in \mathbb{N}_+ : \hat{p}(x) = 0\}.$$

The probability of this (random) event is denoted by the *missing mass*:

$$M_n(X_1, \cdots, X_n) := p(E_n) = \sum_{x \in \mathbb{N}_+} p(x) \mathbf{1}\{\hat{p}(x) = 0\}.$$

The question we strive to answer in this paper is: "Can we learn the missing mass when p is an arbitrary distribution on $\mathbb{N}_+$?" Definition 1 phrases this precisely in the learning framework.

Definition 1. *An* estimator *is a sequence of functions* $\hat{M}_n(x_1, \cdots, x_n) : \mathbb{N}_+^n \to [0,1]$. *We say that an estimator* learns *the missing mass in relative error with respect to a family* $\mathcal{P}$ *of distributions, if for every* $p \in \mathcal{P}$ *and every* $\epsilon, \delta > 0$ *there exists* $n_0(p, \epsilon, \delta)$ *such that for all* $n > n_0(p, \epsilon, \delta)$*:*

$$\mathbf{P}_p \left\{ \left| \frac{\hat{M}_n(X_1, \cdots, X_n)}{M_n(X_1, \cdots, X_n)} - 1 \right| < \epsilon \right\} > 1 - \delta.$$

The learning is said to be distribution-free, *if* $\mathcal{P}$ *consists of* all *distributions on* $\mathbb{N}_+$.

In this framework, our question becomes whether we can distribution-free learn the missing mass in relative error. It is obvious that the empirical estimator $\hat{p}(E_n)$ gives us the trivial answer of 0, and cannot learn the missing mass. A popular alternative is the Good–Turing estimator of the missing mass, which is the fraction of singletons in the data:

$$G_n := \sum_{x \in \mathbb{N}_+} \frac{1}{n} \mathbf{1}\{n\hat{p}(x) = 1\}.$$

The Good–Turing estimator has many interpretations. Its original derivation by Good [1] uses an empirical-Bayes perspective. It can also be thought of as a leave-one-out cross-validation estimator, which contributes to the missing set if and only if the holdout appears exactly once in the data. Fundamentally, G_n derives its form and its various properties from the simple fact that:

$$\mathbf{E}[G_n] = \sum_{x \in \mathbb{N}_+} p(x)(1 - p(x))^{n-1} = \mathbf{E}[M_{n-1}].$$

A study of G_n in the learning framework was first undertaken by McAllester and Schapire [2] and continued later by McAllester and Ortiz [3]. Some further refinement and insight was also given later by Berend and Kontorovich [4]. These works focused on additive error. Ohannessian and Dahleh [5] shifted the attention to relative error, establishing the learning property of the Good–Turing estimator with respect to the family of heavy-tailed (roughly power-law) distributions, e.g., $p(x) \propto x^{-1/\alpha}$ with $\alpha \in (0,1)$. This work also showed that Good–Turing *fails* to learn the missing mass for geometric distributions, and therefore does not achieve distribution-free learning. More recently, Ben-Hamou et al. [6] provide a comprehensive and tight set of concentration inequalities, which can be interpreted in the current framework, and which further demonstrate that Good–Turing can learn with respect to heavier-than-geometric light tails, e.g. the family that includes $p(x) \propto 2^{-x^\gamma}$ with $\gamma \in (0,1)$ (see the definition in Section 4.3 and Remark 4.3 in that paper), in addition to power-laws.

These results leave open the important question: does there exist some *other* estimator that can learn the missing mass in relative error for *any* distribution p? Our contributions are:

- We prove that there are no such estimators, thus providing the first such "no free lunch" theorem for learning about rare events. The first insight to glean from this impossibility result is that one is justified to use further structural assumptions. Furthermore, the proof relies on an implicit construction that uses a dithered geometric distribution. In doing so, it shows that the failure of the Good–Turing estimator for light-tailed distributions is not a weakness of the procedure, but is rather due to a fundamental barrier. Conversely, the success of Good–Turing for heavier-than-geometric and power laws shows its universality, in some restricted sense. In particular, in concrete support to folklore (e.g., [7]), we can state that for estimating probabilities of rare events, heavy tails are both necessary and sufficient.
- We extend this result to continuous tail estimation.
- We show, on a positive note, that upon restricting to parametric light-tailed families learning may be possible. In particular, we show that for the geometric family the natural plug-in estimator learns the missing mass in relative error. As an ancillary result, we prove an instance-by-instance convergence rate, which can be interpreted as a weak sample complexity. For this, we establish

some sharp concentration results for the gaps in geometric distributions, which may be of independent interest.

The paper is organized as follows. In Section 2, we present our main result, with a detailed exposition of the proof. In Section 3 we discuss questions of weak versus strong learnability, we give an immediate extension to continuous tail estimation, show that parametric light-tailed learning is possible, comment further on the Good–Turing estimator, and concisely place this result in the context of a chief motivating application, that of computational linguistics. Lastly, we conclude in Section 4 with a summary and open questions.

2. Main Result

Our main result is stated as follows. The rest of this section is dedicated to its detailed proof.

Theorem 1. *There exists a positive $\epsilon > 0$ and a strictly increasing sequence $(n_k)_{k=1,2,\dots}$, such that for every estimator $\hat{M}_n$ there exists a distribution $p^\star$, such that for all k:*

$$\mathbf{P}_{p^\star}\left\{\left|\frac{\hat{M}_{n_k}}{M_{n_k}} - 1\right| > \epsilon\right\} > \epsilon. \tag{1}$$

In particular, it follows that it is impossible to perform distribution-free learning of the missing mass in relative error.

Remark 1. *Our proof below implies the statement of the theorem with $\epsilon = 10^{-4}$ and $n_k = 6.5 \times 2^k$, but we did not make an honest effort to optimize these parameters.*

2.1. Proof Outline

Consider the family $\mathcal{P}_{\beta,m}$ of β, m-dithered geometric$(\frac{1}{2})$ distributions, where the mass of each outcome beyond a symbol m of a geometric$(\frac{1}{2})$ random variable is divided between two symbols, with a fraction β in one and $1-\beta$ in the other. The individual distributions in this family differ only by which of each pair of such symbols gets which fraction. More precisely:

Definition 2. *The β, m-dithered geometric$(\frac{1}{2})$ family, for a given choice of $\beta \in (0, \frac{1}{2})$ and $m \in \mathbb{N}_+$, is a collection of distributions parametrized by the dithering choices $\theta \in \{\beta, 1-\beta\}^{\mathbb{N}_+}$, $\theta := (\theta_1, \theta_2, \cdots, \theta_j, \cdots)$, as follows:*

$$\mathcal{P}_{\beta,m} = \Bigg\{ p_\theta : p_\theta(x) = \frac{1}{2^x}, x = 1, \cdots, m; \\ p_\theta(m+2j-1) = \frac{\theta_j}{2^{m+j}},\ p_\theta(m+2j) = \frac{1-\theta_j}{2^{m+j}},\ j \in \mathbb{N}_+,\ \theta \in \{\beta, 1-\beta\}^{\mathbb{N}_+} \Bigg\}. \tag{2}$$

The intuition of the proof of Theorem 1 is that within such light-tailed families, two distributions may have very similar samples and thus estimated values, yet have significantly different true values of the missing mass. This follows the general methodology of many statistical lower bounds. We now state the outline of the proof. We choose a subsequence of the form $n_k = C2^k$. We set $\beta = 1/4$, $m = 1$, and $C = 6.5$. The value of $\epsilon > 0$ is made explicit in the proof, and depends only on these choices. We proceed by induction.

- We show that there exists $\theta_1^\star$ such that for all θ with $\theta_1 = \theta_1^\star$ we have for $n = n_1$:

$$\mathbf{P}_{p_\theta}\left\{\left|\frac{\hat{M}_n}{M_n} - 1\right| > \epsilon\right\} > \epsilon. \tag{3}$$

- Then, at every step $k > 1$:

(H) We start with $(\theta_1^\star, \cdots, \theta_{k-1}^\star)$ such that for all θ with $(\theta_1, \cdots, \theta_{k-1}) = (\theta_1^\star, \cdots, \theta_{k-1}^\star)$, Inequality (3) holds for $n = n_1, \cdots, n_{k-1}$.

(*) We then show that it must be that for at least one of $\tilde{\theta} = \beta$ or $\tilde{\theta} = 1 - \beta$, for all θ with $(\theta_1, \cdots, \theta_k) = (\theta_1^\star, \cdots, \theta_{k-1}^\star, \tilde{\theta})$, Inequality (3) holds additionally for $n = n_k$. We select $\theta_k^\star$ to be the corresponding $\tilde{\theta}$.

- This induction produces an infinite sequence $\theta^\star \in \{\beta, 1-\beta\}^{\mathbb{N}_+}$, and the desired distribution in Theorem 1 can be chosen as $p^\star = p_{\theta^\star}$, since it is readily seen to satisfy the claim for each n_k, by construction.

2.2. Proof Details

We skip the proof of the base case, since it is mostly identical to that of the induction step. Therefore, in what follows we assume that $(\theta_1^\star, \cdots, \theta_{k-1}^\star)$ satisfies the inductive hypothesis (H), and we would like to prove that the selection in (*) can always be done. Let us denote the two choices of parameters by

$$\theta := (\theta_1^\star, \cdots, \theta_{k-1}^\star, \beta, \theta_{k+1}, \cdots),$$

and

$$\theta' := (\theta_1^\star, \cdots, \theta_{k-1}^\star, 1-\beta, \theta'_{k+1}, \cdots),$$

and let us refer to $(\theta_{k+1}, \cdots)$ and $(\theta'_{k+1}, \cdots)$ by the *trailing parameters*. What we show in the remainder of the proof is that with two arbitrary sets of trailing parameters, we cannot have two simultaneous violations of Inequality (3) (for both θ and θ'). That is, we cannot have both:

$$\mathbf{P}_{p_\theta}\left\{\left|\frac{\hat{M}_{n_k}}{M_{n_k}} - 1\right| > \epsilon\right\} < \epsilon \quad \text{and} \quad \mathbf{P}_{p_{\theta'}}\left\{\left|\frac{\hat{M}_{n_k}}{M_{n_k}} - 1\right| > \epsilon\right\} < \epsilon. \tag{4}$$

This is stated in Lemma 3, in the last portion of this section. To see why this is sufficient to show that the selection in (*) can be done, consider first the case that Inequality (3) with $n = n_k$ is upheld for both θ and θ' with any two sets of trailing parameters. In this case we can arbitrarily choose $\theta_k^\star$ to be either β or $1-\beta$, since the induction step is satisfied. We can therefore focus on the case in which this fails. That is, for either θ or θ' a choice of trailing parameters can be made such that Inequality (3) with $n = n_k$ is *not* satisfied, and therefore one of the two cases in (4) holds [say, for example, for θ]. Fix the corresponding trailing parameters [in this example, $(\theta_{k+1}, \cdots)$]. Then, for *any* choice of the *other* set of trailing parameters [in this example, $(\theta'_{k+1}, \cdots)$], Lemma 3 precludes a violation of Inequality (3) for $n = n_k$ by the other choice [in this example, θ']. Therefore this choice can be selected for θ_k [in this example, $\theta_k = 1 - \beta$.]

By using the *coupling* device and restricting ourselves to a *pivotal event*, we formalize the aforementioned intuition that the estimator may not distinguish between two separated missing mass values, and deduce that both statements in (4) cannot hold simultaneously.

2.2.1. Coupling

Definition 3. *A* coupling *between two distributions p and p' on $\mathbb{N}_+$ is a joint distribution q on $\mathbb{N}_+^2$, such that the first and second marginal distributions of q revert back to p and p' respectively.*

Couplings are useful because probabilities of events on each side may be evaluated on the joint probability space, while forcing events of interest to occur in an orchestrated fashion. Going back to

our induction step and the specific choices θ and θ' with arbitrary trailing parameters, we perform the following coupling. For $(x, x') \in \mathbb{N}_+^2$, let

$$q(x,x') = \begin{cases} p_\theta(x) = p_{\theta'}(x') & ; \text{ if } x = x' < m+2k-1; \\ \beta/2^{m+k} & ; \text{ if } x = x' = m+2k-1, \text{or if } x = x' = m+2k, \\ (1-2\beta)/2^{m+k} & ; \text{ if } x = m+2k \text{ and } x' = m+2k-1; \\ p_\theta(x)p_{\theta'}(x')/2^{m+k} & ; \text{ if } x, x' > m+2k; \\ 0 & ; \text{ otherwise.} \end{cases} \tag{5}$$

It is easy to verify that q in Equation (5) is a coupling between p_θ and $p_{\theta'}$ as in Definition 3. Now let us observe the consequences of this choice. If X, X' are generated according to q, then if either is in $\{1, \cdots, m+2k-2\}$ then both values are *identical*. If either is in $\{m+2k+1, \cdots\}$ then so is the other, but otherwise the two values are conditionally independent. If either is in $\{m+2k-1, m+2k\}$, so is the other, and the conditional probability is given by:

x \ x'	$m+2k-1$	$m+2k$
$m+2k-1$	β	0
$m+2k$	$1-2\beta$	β

Now consider coupled data $(X_i, X_i')_{i=1,\cdots,n}$ generated as i.i.d. samples from q. It follows that, marginally, the X-sequence is i.i.d. from p_θ, and so is the X'-sequence from $p_{\theta'}$. Any event B that is exclusively X-measurable or B' that is exclusively X'-measurable has the same probability under the coupled measure. That is,

$$\mathbf{P}_{p_\theta}(B) = \mathbf{P}_q(B) := q^n(B \times \mathbb{N}_+^n)$$

and

$$\mathbf{P}_{p_{\theta'}}(B') = \mathbf{P}_q(B') := q^n(\mathbb{N}_+^n \times B').$$

In what follows we work only with coupled data, and use simply the shorthand $\mathbf{P}$ to mean $\mathbf{P}_q$.

2.2.2. Pivotal Event

The event we would like to work under is that of the coupled samples being identical, while exactly covering the range $1, \cdots, m+2k-1$. Let's call this the *pivotal event* and denote it by:

$$A_k := \bigcap_{i=1}^{n_k} \{X_i = X_i'\} \ \cap \ \Big\{\{X_1, \ldots, X_{n_k}\} = \{1, \cdots, m+2k-1\}\Big\}. \tag{6}$$

The reason A_k interests us is that it encapsulates the aforementioned intuition.

Lemma 1. *Under event A_k, the coupled missing masses are distinctly separated,*

$$\frac{M_{n_k}}{M'_{n_k}} = \frac{2-\beta}{1+\beta},$$

while any estimator cannot distinguish the coupled samples,

$$\hat{M}_{n_k} = \hat{M}'_{n_k}.$$

Proof. The confusion of any estimator is simply due to the fact that under A_k, the coupling forces all samples to be identical $X_i = X_i'$, for all $i = 1, \cdots, n_k$.

Thus $\hat{M}_{n_k} = \hat{M}'_{n_k}$, since estimators only depend on the samples and not the probabilities.

The missing masses, on the other hand, do depend on both the samples and the probabilities and thus they differ. But the event A_k makes the set of missing symbols simply the tail $m+2k, m+2k+1, \cdots$, so we can compute the missing masses exactly:

$$M_{n_k} = p_\theta(m+2k) + \sum_{x=m+2k+1}^{\infty} p_\theta(x) = \frac{1-\theta_k}{2^{m+k}} + \frac{1}{2^{m+k}} = (2-\beta)2^{-m-k}, \text{ and}$$

$$M'_{n_k} = p_{\theta'}(m+2k) + \sum_{x=m+2k+1}^{\infty} p_{\theta'}(x) = \frac{1-\theta'_k}{2^{m+k}} + \frac{1}{2^{m+k}} = (1+\beta)2^{-m-k},$$

where $\frac{1}{2^{m+k}}$ follows from the usual geometric sum. The claim follows. □

We now show that A_k has always a positive probability, bounded away from zero.

Lemma 2. *For $\beta = 1/4$, $m = 1$, $C = 6.5$ and $n_k = C2^k$, there exists a positive constant $\eta > 0$ that does not depend on θ, such that for all k, $\mathbf{P}(A_k) > \eta$. We can explicitly set $\eta = 2 \cdot 10^{-4}$.*

Proof. Please note that A_k in Equation (6) overspecifies the event. In fact, only forcing the exact coverage of $1, \cdots, m+2k-1$ is sufficient, since this implies in turn that the coupled samples are identical. Recalling the coupling of Equation (5), this is immediate for symbols in $1, \cdots, m+2k-2$, and follows for $m+2k-1$ since $m+2k$ is not allowed in this event. We can then write $A_k = A_{k,1} \cap A_{k,2}$, dividing the exact coverage to the localization in the range and the representation of each symbol by at least one sample:

$$\begin{aligned} A_{k,1} &= \left\{\bigcup_{i=1}^{n_k}\{X_i\} \subseteq \{1, \cdots, m+2k-1\}\right\} && \text{(localization)}, \\ A_{k,2} &= \left\{\bigcup_{i=1}^{n_k}\{X_i\} \supseteq \{1, \cdots, m+2k-1\}\right\} && \text{(representation)}. \end{aligned}$$

Let α be the probability of (X_i, X'_i) for a given i being in $\{(1,1), \cdots, (m+2k-1, m+2k-1)\}$. From the coupling in Equation (5) and the structure of the dithered family in Equation (2), we see that for up to $m+2k-2$ this probability sums up to the $m+k-1$ first terms of a geometric$(\frac{1}{2})$, and for $(m+2k-1, m+2k-1)$ the coupling assigns it $\beta/2^{m+k}$, thus:

$$\alpha = \sum_{x=1}^{m+2k-1} q(x,x) = 1 - \frac{1}{2^{m+k-1}} + \frac{\beta}{2^{m+k}}.$$

We can then explicitly compute:

$$\mathbf{P}(A_{k,1}) = \alpha^{n_k} = \left(1 - \frac{1}{2^{m+k-1}} + \frac{\beta}{2^{m+k}}\right)^{n_k} =: \eta_1(k).$$

Meanwhile, the complement of $A_{k,2}$ is the event that at least one of $\{(1,1), \cdots, (m+2k-1, m+2k-1)\}$ does not appear, that is $A_{k,2}^c = \bigcup_{x=1}^{m+2k-1}\{x \notin \cup\{X_i\}\}$. Conditionally on $A_{k,1}$, the occurrence probabilities of these symbols are simply normalized by α, that is $\mathbf{P}(x \notin \cup X_i | A_{k,1}) = [1 - q(x,x)/\alpha]^{n_k}$. Thus, by a union bound, we have:

$$\begin{aligned}
\mathbf{P}(A_{k,2}|A_{k,1}) &= 1-\mathbf{P}(A_{k,2}^{\mathsf{c}}|A_{k,1}) \\
&\geq 1-\textstyle\sum_{x=1}^{m+2k-1}\mathbf{P}(x\notin\cup X_i|A_{k,1}) \\
&= 1-\textstyle\sum_{x=1}^{m+2k-1}\left[1-q(x,x)/\alpha\right]^{n_k} \\
&\geq 1-\textstyle\sum_{x=1}^{m+2k-1}\left[1-q(x,x)\right]^{n_k} \\
&= 1-\textstyle\sum_{x=1}^{m}\left(1-\frac{1}{2^x}\right)^{n_k} \\
&\quad -\textstyle\sum_{j=1}^{k-1}\left[\left(1-\frac{\beta}{2^{m+j}}\right)^{n_k}+\left(1-\frac{1-\beta}{2^{m+j}}\right)^{n_k}\right]-\left(1-\frac{\beta}{2^{m+k}}\right)^{n_k} \\
&\geq 1-\textstyle\sum_{x=1}^{m}\left(1-\frac{1}{2^x}\right)^{n_k}-2\sum_{j=1}^{k-1}\left(1-\frac{\beta}{2^{m+j}}\right)^{n_k}-\left(1-\frac{\beta}{2^{m+k}}\right)^{n_k} =: \eta_2(k),
\end{aligned}$$

where the last inequality follows from the fact that $\beta < \frac{1}{2}$. Therefore,

$$\mathbf{P}(A_k) = \mathbf{P}(A_{k,1}\cap A_{k,2}) = \mathbf{P}(A_{k,1})\mathbf{P}(A_{k,2}|A_{k,1}) \geq \eta_1(k)\eta_2(k) \geq \inf_{k\geq 1}\eta_1(k)\eta_2(k) =: \eta.$$

We now use our choices of $\beta = 1/4$, $m = 1$, $C = 6.5$, and $n_k = C2^k$, to bound this worst-case η. In particular, we can verify that $\eta \geq 2\cdot 10^{-4}$, and it follows as claimed that the pivotal event has always a probability bounded away from zero. □

2.2.3. Induction Step

We now combine all the elements presented thus far to complete the proof of Theorem 1 by establishing the following claim, which we have shown in the beginning of the detailed proof section to be sufficient for the validity of the induction step. In particular, we restate Equation (4) under the coupling of Equation (5).

Lemma 3. *Let*

$$\theta := (\theta_1^\star,\cdots,\theta_{k-1}^\star,\beta,\theta_{k+1},\cdots), \text{ and } \theta' := (\theta_1^\star,\cdots,\theta_{k-1}^\star,1-\beta,\theta'_{k+1},\cdots),$$

with arbitrary *trailing parameters* $(\theta_{k+1},\cdots)$ *and* $(\theta'_{k+1},\cdots)$. *Let* q *be the coupling of Equation* (5), *and let* $B_k = \left\{\left|\hat{M}_{n_k}/M_{n_k}-1\right|>\epsilon\right\}$ *and* $B'_k = \left\{\left|\hat{M}'_{n_k}/M'_{n_k}-1\right|>\epsilon\right\}$. *Then given our choices of* $\beta = 1/4$, $m = 1$, $C = 6.5$ *and* $n_k = C2^k$, *if* $\epsilon < 10^{-4}$ *we cannot simultaneously have*

$$\mathbf{P}_q(B_k) < \epsilon \text{ and } \mathbf{P}_q(B'_k) < \epsilon.$$

Proof. Please note that this choice of ϵ means that $\epsilon < \eta/2$, where η is as in Lemma 2. Recall the pivotal event A_k, and assume, for the sake of contradiction, that both probability bounds $\mathbf{P}(B_k) < \epsilon$ and $\mathbf{P}(B'_k) < \epsilon$ hold. Please note that if B_k^{c} holds, it means that

$$\hat{M}_{n_k}/M_{n_k} \in (1-\epsilon, 1+\epsilon), \tag{7}$$

and similarly if $B_k'^{\mathsf{c}}$ holds, it means that

$$\hat{M}'_{n_k}/M'_{n_k} \in (1-\epsilon, 1+\epsilon). \tag{8}$$

By making our hypothesis, we are asserting that these events have high probabilities, $1-\epsilon$, under both p_θ and $p_{\theta'}$ distributions, and that thus the estimator is effectively $(1\pm\epsilon)$-close to the true value of the missing mass. Yet, we know that this would be violated under the pivotal event, which occurs with positive probability. We now formalize this contradiction.

By Lemma 2, we have that:

$$\left.\begin{aligned}\mathbf{P}(B_k|A_k) = \frac{\mathbf{P}(A_k \cap B_k)}{\mathbf{P}(A_k)} \leq \frac{\mathbf{P}(B_k)}{\mathbf{P}(A_k)} \leq \frac{\epsilon}{\eta} \\ \mathbf{P}(B'_k|A_k) = \frac{\mathbf{P}(A_k \cap B'_k)}{\mathbf{P}(A_k)} \leq \frac{\mathbf{P}(B'_k)}{\mathbf{P}(A_k)} \leq \frac{\epsilon}{\eta}\end{aligned}\right\} \Rightarrow \quad \mathbf{P}(B_k^{\mathsf{c}} \cap B_k'^{\mathsf{c}}|A_k) \geq 1 - 2\frac{\epsilon}{\eta} > 0, \tag{9}$$

where the last inequality is strict, by the choice of $\epsilon < \eta/2$.

On the other hand, recall that by Lemma 1 under A_k we have:

$$\hat{M}_{n_k} = \hat{M}'_{n_k} \quad \text{and} \quad \frac{M_{n_k}}{M'_{n_k}} = \frac{2-\beta}{1+\beta} = \tfrac{7}{5}.$$

By combining this with Equations (7) and (8), we can now see that if $\frac{1+\epsilon}{1-\epsilon} < \frac{7}{5}$, which is satisfied by any choice of $\epsilon < 1/6$, in particular ours, then if B_k^{c} occurs, then B'_k occurs, and conversely if $B_k'^{\mathsf{c}}$ occurs then B_k occurs. For example, say B_k^{c} occurs, then $\hat{M}_{n_k}/M_{n_k} < (1+\epsilon)$:

$$\frac{\hat{M}'_{n_k}}{M'_{n_k}} = \frac{\hat{M}_{n_k}}{\frac{7}{5}M_{n_k}} < \tfrac{5}{7}(1+\epsilon) < 1-\epsilon,$$

implying that Equation (8) is not satisfied, thus B'_k occurs. The end result is that under event A_k, B_k^{c} and $B_k'^{\mathsf{c}}$ cannot occur at the same time, and thus:

$$\mathbf{P}(B_k^{\mathsf{c}} \cap B_k'^{\mathsf{c}}|A_k) = 0.$$

This contradicts the bound in (9), and establishes the lemma. □

3. Discussions

3.1. Weak Versus Strong Distribution-Free Learning

Arguably, a more common notion of learning is a strong version of Definition 1, where the sample complexity is a function of the distribution class rather than the instance. Formally:

Definition 4. *We say that an estimator* learns *the missing mass in relative error* strongly *with respect to a family $\mathcal{P}$ of distributions, if for every $\epsilon > 0, \delta \in (0,1)$ there exists $n_0(\mathcal{P}, \epsilon, \delta)$ such that for all $p \in \mathcal{P}$ and all $n > n_0(\mathcal{P}, \epsilon, \delta)$:*

$$\mathbf{P}_p\left\{\left|\frac{\hat{M}_n(X_1, \cdots, X_n)}{M_n(X_1, \cdots, X_n)} - 1\right| < \epsilon\right\} > 1 - \delta.$$

The learning is said to be strongly distribution-free, *if $\mathcal{P}$ consists of* all *distributions on $\mathbb{N}_+$.*

The distinction here is similar to that of uniform versus pointwise convergence. Clearly, the existence of a strong learner implies the existence of a weak learner. Conversely, as we have shown that there is no weakly distribution-free learner, there is also no strongly distribution-free learner. However, the ability to choose a different distribution at every sample size n makes it very easy to show this corollary directly. For example, we can consider two distributions p and q with $p(1) = 1 - \frac{1}{n^2}$ and $q(1) = 1 - \frac{1}{100n^2}$, both of which would result with overwhelming probability in a length-n sequence consisting entirely of this first symbol. Thus any estimator would need to predict the same missing mass with high probability. However, the rest of the symbols would have probabilities differing by a factor of 100 between the two models, and thus any estimator would be misguided for at least one of the two cases.

The relevance of the current contribution is rooted in the plausible yet misguided optimism that although we may not do well in such a worst-case paradigm, there is more hope if we first fix the instance and then study asymptotics. Our "no free lunch" theorem indeed shows the more subtle fact that there are always bad instances for every estimator, and thus even such weak learning is fundamentally impossible.

Such a contrast between weak/strong learning has also been appreciated in the classical learning literature, notably in the work of Antos and Lugosi [8]. The notions there are framed in the negative, which is why the weak/strong terminology is reversed. A traditional minimax lower bound in that context states that for any sequence of concept learners $\hat{g}_n$ at each n we can find a distribution for which the expected cumulative classification error is lower bounded by the complexity of the concept class. Analogously, *not* being able to strong learn as in Definition 4 means that for any estimator $\hat{M}_n$ at each n we can find a distribution for which the relative error stays away from zero. By demanding a *strong* performance from a learner/estimator, we are able to give only a *weak* guarantee. In particular, it could be too loose for a fixed distribution that doesn't vary with n. [8] contributes by giving lower bounds that hold infinitely often for any sequence of concept learners $\hat{g}_n$ but for a distribution choice that is adversarial in advance, fixed for all n. Analogously, *not* being able to (weak) learn as in Definition 1 means that for any estimator $\hat{M}_n$ there exists a distribution, fixed for all n, for which the relative error stays away from zero for infinitely many n. The lower bounds of Antos and Lugosi [8] can now be tighter, which is why they call their results *strong* minimax lower bounds. In the context of the present paper, of course, the lower bounds correspond to the impossibility result, which is thus stronger since it doesn't even hold for a fixed distribution.

3.2. Generalization to Continuous Tails

A closely related problem to learning the missing mass is that of estimating the tail of a probability distribution. In the simplest setting, the data consists of $Y_1, \cdots, Y_n$ that are i.i.d. samples from a continuous distribution on $\mathbb{R}$. Let F be the cumulative distribution function. The task in question is that of estimating the tail probability

$$W_n = 1 - F\left(\max_{i=1}^{n} Y_i\right),$$

that is the probability that a new sample exceeds the maximum of all samples seen in the data.

One can immediately see the similarity with the missing mass problem, as both problems concern estimating probabilities of underrepresented events. We can use essentially the same learning framework given by Definition 1, and prove a completely parallel impossibility result.

Theorem 2. *There exists $\epsilon > 0$ and a subsequence $(n_k)_{k=1,2,\ldots}$, such that for every estimator $\hat{W}_n$ of W_n there exists a continuous distribution $F^\star$, such that for all k:*

$$\mathbf{P}_{F^\star}\left\{\left|\frac{\hat{W}_{n_k}}{W_{n_k}} - 1\right| > \epsilon\right\} > \epsilon.$$

In particular, it follows that it is impossible to perform distribution-free learning of the tail probability in relative error.

Proof. (Sketch) The discrete version of this theorem is a trivial extension of Theorem 1, since in the proof of the latter the pivotal event forced the missing mass to be a tail probability. The potential strengthening of Theorem 2 comes from insisting on a continuous $F^\star$. The same techniques may be adapted in this case, such as by dithering an exponential distribution, where a base exponential density is divided into intervals, and mass is moved between pairs of adjacent intervals by scaling the density the same way as β dithers the geometric. The adversarial distribution for a given estimator can then be

chosen from this family. In order not to repeat the same aguments, however, we instead prove this result via a reduction. The details can be found in the Appendix A. Namely, we show that discrete tail estimation can be reduced to continuous tail estimation. Since the former is impossible, so is the latter. □

Theorem 2 gives a concrete justification of why it is important to make regularity assumptions when extrapolating distribution tails. This is of course the common practice of extreme value theory, see, for example [9]. Some impossibility results concerning the even more challenging problem of estimating the density of the maximum were already known [10], but to the best of our knowledge this is the first result asserting it for tail probability estimation as well.

3.3. Learning in Various Families

Ben-Hamou et al. [6] (Corollary 5.3) gives a very clean characterization of a sufficient learnable family, which encompasses the one covered by Ohannessian and Dahleh [5].

Theorem 3 ([6])**.** *Denote the expected number of single-occurrence and double-occurrence symbols by $\Phi_{n,1}$ and $\Phi_{n,2}$ respectively:*

$$\Phi_{n,1} := \mathbf{E}\left[\sum_{x\in\mathbb{N}_+} \mathbf{1}\{n\hat{p}_n(x) = 1\}\right] = \sum_{x\in\mathbb{N}_+} np(x)[1-p(x)]^{n-1}, \text{ and}$$

$$\Phi_{n,2} := \mathbf{E}\left[\sum_{x\in\mathbb{N}_+} \mathbf{1}\{n\hat{p}_n(x) = 2\}\right] = \sum_{x\in\mathbb{N}_+} \frac{n(n-1)}{2} p(x)^2[1-p(x)]^{n-2}.$$

Let $\mathcal{H}$ be the family defined by:

$$\mathcal{H} := \left\{p \,:\, \Phi_{n,1} \to \infty \text{ and } \frac{\Phi_{n,2}}{\Phi_{n,1}} \text{ remains bounded as } n \to \infty\right\}.$$

The Good–Turing estimator learns the missing mass in relative error with respect to $\mathcal{H}$.

The proof relies on power moment concentration inequalities (such as Chebyshev's). The $\Phi_{n,1} \to \infty$ property embodies the heavy-tailed nature, since it says that rare events occur often. The condition that $\frac{\Phi_{n,2}}{\Phi_{n,1}}$ (i.e., its lim sup) remains bounded is a smoothness condition, since $\Phi_{n,2}$ roughly captures the variance of the number of singletons (see [6], Proposition 3.3). For us, this is instructive because one could readily verify that the condition of Theorem 3 fails for geometric (and dithered geometric) distributions. We can thus see that in some sense Good–Turing captures a maximal family of learnable distributions. In particular, we now know that the complement of $\mathcal{H}$ is not learnable.

Considering how sparse the dithered geometric family is, the failure of any estimator to learn the missing mass with respect to it may seem discouraging. (Please note that Theorem 1 holds even if the estimator *is aware* that this is the class it is paired with.) However, if we restrict ourselves to smooth parametric families within light tails then the outlook can be brighter. We illustrate this with the case of the geometric family.

Theorem 4. *Let $\mathcal{G}$ be the class of geometric distributions, parametrized by $\alpha \in (0,1)$:*

$$p_\alpha(x) = (1-\alpha)\alpha^{x-1}, \qquad \text{for } x \in \mathbb{N}_+.$$

Let $\hat{\alpha}_n = 1 - \frac{n}{\sum X_i}$ be the maximum likelihood estimator of the parameter, and define the plug-in estimator:

$$\check{M}_n = \sum_{x\in\mathbb{N}_+} (1-\hat{\alpha}_n)\hat{\alpha}_n^x \mathbf{1}\{n\hat{p}_n(x) = 0\}$$

Then $\check{M}_n$ learns the missing mass in relative error with respect to $\mathcal{G}$.

Proof. (Sketch) The proof consists of pushing forward the convergence of the parameter to that of the entire distribution using continuity arguments, and then specializing to the missing mass. The details can be found in the Appendix B. □

3.4. Learning the Missing Mass in Additive Error and Learning Other Related Quantities

As mentioned in the introduction, a good part of the work on learning the missing mass focused on additive error [2–4]. Recently, minimax lower bounds were given for the additive error in [11] and [12]. Note however that relative error bounds cannot be deduced from these (nor any other way, given the impossibility established here.) A related problem to learning the missing mass in relative error is that of learning a distribution in KL-divergence loss. This averages all log-relative errors (missing or otherwise). This averaging scales the log-relative errors by the rare probability and attenuates the kind of gaps discussed in our present context. One thus hopes to have more optimistic results. Indeed, the Good–Turing estimator was recently shown to be adaptive/competitive for distribution learning in KL-divergence [13]. A similar result in the context of distribution learning in total variation was given in [14]. In the language of Section 3.2, being competitive can be understood as an intermediate characterization between weak and strong learning. Lastly, one could be interested in learning other properties of distributions that are intimately related to the rare component. Entropy and support size are two of these. In [15] a traditional minimax bound was established for these quantities, which was then further distilled in [16]. Another related problem is predicting the growth of the support as more observations are made. This was characterized very precisely in [17], where one can find further pointers on this very old problem. Some of these results may give the impression that nothing further can be gained from structural assumptions. However, rates can generally be refined whenever such structure exists. See for example [18] for refined competitive rates in distribution estimation and [19] for similar results in the predictive/compression setting. These results use tail characterizations, similar to those in extreme value theory [10].

3.5. N-Gram Models and Bayesian Perspectives

One of the prominent applications of estimating the missing mass has been to computational linguistics. In that context, it is known as *smoothing* and is used to estimate N-gram transition probabilities. The importance of accurately estimating the missing mass, and in particular in a relative-error sense, comes from the fact that N-grams are used to score test sentences using log-likeliehoods. Test sentences often have transitions that are never seen in the training corpus, and thus in order for the inferred log-likelihoods to accurately track the true log-likelihood, these rare transitions need to be assigned meaningful values, ideally as close to the truth as possible. As such, various forms of smoothing, including Good–Turing esimation, have become an essential ingredient of many practical algorithms, such as the popular method proposed by Kneser and Ney [20].

In the context of N-gram learning, a separate Bayesian perspective was also proposed. One of the earliest to introduce this were [21] using a Dirichlet prior. This was shown to not be very effective, and we now understand that it is due to the fact that (1) the Dirichlet process produces light tails while language is often heavy-tailed and, even if it were; (2) rare probabilities are hard to learn for large light-tailed families. The natural progression of these Bayesian models led to the use of the two-parameter Poisson-Dirichlet prior [22], which was suggested initially by [23]. Despite employing sophisticated inference techniques, the missing mass estimator that resulted from these models closely followed the Good–Turing estimator (for a sharp analysis of this correspondence, see Falahatgar et al. [18].) In light of the present work, this is not surprising since the two-parameter Poisson-Dirichlet process almost surely produces heavy-tailed distributions, and any two algorithms that learn the missing mass are bound to have the same qualitative behavior.

4. Summary

In this paper, we have considered the problem of learning the missing mass, which is the probability of all unseen symbols in an i.i.d. draw from an unknown discrete distribution. We have phrased this in the probabilistic framework of learning. Our main contribution was to show that it is not possible to learn the missing mass in a completely distribution-free fashion.

In other words, no single estimator can do well for all distributions. We have given a detailed account of the proof, emphasizing the intuition of how failure can occur in large light-tailed families. We have also placed this work in a greater context, through some discussions and extensions of the impossibility result to continuous tail probability estimation, and by showing that smaller, parametric, light-tailed families may be learnable.

An initial impetus for this paper and its core message is that assuming further structure can be necessary in order to learn rare events. Further structure, of course, is nothing more than a form of regularization. This is a familiar notion to the computational learning community, but for a long time the Good–Turing estimator enjoyed favorable analysis that focused on additive error, and evaded this kind of treatment. The essential ill-posedness of the problem was uncovered by studying relative error. But lower bounds cannot be deduced from the failure of particular algorithms. Our result thus completes the story, and we can now shift our attention to studying the landscape that is revealed.

The most basic set of open problems concerns establishing families that allow learning of the missing mass. We have seen in this paper some such families, including the heavy-tailed family learnable by the Good–Turing estimator, and simple smooth parametric families, learnable using plug-in estimators. How do we characterize such families more generally? The next layer of questions concerns establishing convergence rates, i.e., strong learnability, via both lower and upper bounds. The fact that a family of distributions allows learning does not mean that such rates can be established. This is because any estimator may be faced with arbitrarily slow convergence, by varying the distribution in the family. In other words we may be faced with a lack of uniformity. How do we control the convergence rate? Lastly, when learning is not possible, we may want to establish how gracefully an estimator can be made to fail. Understanding these limitations and accounting for them can be critical to the proper handling of data-scarce learning problems.

Author Contributions: Conceptualization, M.I.O.; methodology, E.M.; formal analysis, M.I.O. and E.M.; writing, M.I.O.

Funding: The first author was supported by NSF grants DMS 1106999 and CCF 1320105, by DOD ONR grant N00014-14-1-0823, and by grant 328025 from the Simons Foundation. The second author was supported by funds from the California Institute for Telecommunications and Information Technology (Calit2) while a postdoctoral researcher hosted by Alon Orlitsky at the University of California, San Diego.

Acknowledgments: The authors are grateful to the Simons Institute at the University of California, Berkeley for hosting the programs that make such collaborations possible.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. This Appendix Presents the Proof of Theorem 2.

Let us first consider the tail estimation problem in the discrete case. W_n is still well-defined here. But in order to clearly delineate the discrete case, let us define/rename the tail probability as:

$$T_n := \sum_{x > \max_i X_i} p(x), \tag{A1}$$

and let $\hat{T}_n$ denote an arbitrary estimator of T_n. It is immediate that T_n is identical to W_n in the discrete case. Next, note that in the proof of Theorem 1, the pivotal event A_k given by Equation 6 forces the missing mass to be a tail probability (see the proof of Lemma 1.) That is, under A_k, $T_{n_k} = M_{n_k}$. It therefore follows that Lemma 1 holds with M_n and $\hat{M}_n$ replaced by T_n and $\hat{T}_n$ respectively. Lemma 2 only depends on the definition of the pivotal event, not the missing mass. And the argument of

Lemma 3 follows exactly identically with again M_n and $\hat{M}_n$ replaced by T_n and $\hat{T}_n$ respectively everywhere in the statement and its proof. Consequently, the proof of Theorem 1 produces the following parallel result. There exist $\gamma > 1$ and a sequence $(n_k)_{k=1,2,\cdots}$ (both universal) such that for every $\hat{T}_n$ we can find a distribution $p^\star$ (which does depend on $\hat{T}_n$), such that for all k (a slight notational change):

$$\mathbf{P}_{p^\star}\left\{\frac{\hat{T}_{n_k}}{T_{n_k}} \notin \left[\gamma^{-1}, \gamma\right]\right\} > \gamma - 1. \tag{A2}$$

So essentially, Theorem 2 is a direct corollary of our main result if we allow any distribution on the real line, including discrete ones. However, we want to show that the same result holds even over the family of properly continuous distributions, which have a density on $\mathbb{R}$ with respect to the Lebesgue measure (if the density is not otherwise restricted). The proof of this fact could be done by paralleling the proof thus far, but by dithering a continuous distribution instead. However, for the sake of novelty, we prove it instead via reduction from discrete to continuous.

We first make a minor generalization of the discrete framework. Observe that nowhere in the proof of Theorem 1 was $\hat{M}_n$ required to be a deterministic function of the observation. So $\hat{M}_n$ and $\hat{T}_n$ could be *randomized*, i.e., depend on an additional random element ζ_n that is independent of the observation and any of the parameters of the problem (see Definition 1.) More rigorously, we include this randomness in the coupling of the proof by simply letting $\zeta_n = \zeta'_n$ always. This way, the samples being identical still implies that the values of the estimators are identical. This is the only property of the estimator we used and thus all the arguments follow.

The rest of the proof basically reduces the ability to learn the discrete tail with a randomized estimator to the ability to learn the continuous tail. The claimed impossibility follows from this reduction. Let us thus assume, for the sake of contradiction, that there exists $\hat{W}_n$ s.t. for all continuous distributions F, for all $\eta > 1$:

$$\mathbf{P}_F\left\{\frac{\hat{W}_n}{W_n} \in \left[\eta^{-1}, \eta\right]\right\} \to 0, \text{ as } n \to \infty. \tag{A3}$$

Next, recall that by construction we more precisely know that $p^\star$ from earlier can be chosen from the $(\beta = \frac{1}{4}, m = 1)$-dithered geometric$(\frac{1}{2})$ family. It is easy to check that for all such distributions, the tail is comparable to its preceding symbol's probability. In particular it is not much smaller: there exists $\kappa < \infty$ such that for all p in this family

$$\max_x \frac{p(x)}{\sum_{x'>x} p(x')} \le \kappa,$$

and we can always choose $\kappa > \sqrt{\gamma} - 1$ (by capping from below). For reasons that will shortly become apparent, choose any m such that

$$\left(1 - \frac{\sqrt{\gamma} - 1}{\kappa}\right)^m < \frac{\gamma - 1}{2}.$$

Now let G be any continuous distribution on $[0, 1]$. Denote by $\overline{G} := 1 - G$ the corresponding tail probability function and let $\overline{G}^{-1}(t) := \inf\{z : \overline{G}(z) \le t\}$ denote its left inverse. Note two properties: $\overline{G}(z) > t$ and $z < \overline{G}^{-1}(t)$ are equivalent, and $\overline{G}(\overline{G}^{-1}(t)) = t$.

Let $Z_{i,\ell}$ be i.i.d. draws from G, for $i = 1, \cdots, n$ and $\ell = 1, \cdots, m$. Let $\zeta_n = (Z_{i,\ell})_{i=1,\cdots,n;\ \ell=1,\cdots,m}$, and for a given observation sequence $x_1, \cdots, x_n$ in $\mathbb{N}_+^n$, define the randomized estimator:

$$\hat{T}_n(x_1, \cdots, x_n, \zeta_n) := \min_{\ell=1}^{m} \hat{W}_n(x_1 + Z_{1,\ell}, \cdots, x_n + Z_{n,\ell}). \tag{A4}$$

For this specific choice of estimator, let $p^\star$ be the distribution that yields the impossibility result in Equation (A2). Let X_i be i.i.d. drawn from $p^\star$, independently of the infinite array $(Z_{i,\ell})$. Observe that then , for each ℓ, the sequence $Y_{i,\ell} := X_i + Z_{i,\ell}$ is i.i.d. distributed according to a *continuous* distribution $F^\star$, and has its own tail probability $W_{n,\ell} := 1 - F^\star(\max_i Y_{i,\ell})$. Also, let the estimate of this tail probability for each ℓ be denoted by $\hat{W}_{n,\ell} := \hat{W}_n(Y_{1,\ell}, \cdots, Y_{n,\ell})$. Accordingly, based on our assumption, the limiting property in Equation (A3) holds and for any $\eta > 1$:

$$\forall \ell \qquad \mathbf{P}\left\{ \frac{\hat{W}_{n,\ell}}{W_{n,\ell}} \notin \left[\eta^{-1}, \eta\right] \right\} \to 0, \text{ as } n \to \infty.$$

Please note that if for all ℓ we have $\eta^{-1} < \frac{a_\ell}{b_\ell} < \eta$ then also $\eta^{-1} < \frac{\min_\ell a_\ell}{\min_\ell b_\ell} < \eta$. Thus:

$$\begin{aligned} \mathbf{P}\left\{ \frac{\min_\ell \hat{W}_{n,\ell}}{\min_\ell W_{n,\ell}} \notin \left[\eta^{-1}, \eta\right] \right\} &\leq \mathbf{P} \bigcup_{\ell=1}^{m} \left\{ \frac{\hat{W}_{n,\ell}}{W_{n,\ell}} \notin \left[\eta^{-1}, \eta\right] \right\} \\ &\leq m\, \mathbf{P}\left\{ \frac{\hat{W}_{n,1}}{W_{n,1}} \notin \left[\eta^{-1}, \eta\right] \right\} \to 0, \text{ as } n \to \infty. \end{aligned} \tag{A5}$$

This means that $\min_\ell \hat{W}_{n,\ell} \equiv \hat{T}_n$ (recall the definition of Equation (A4)) is a good estimator of $\min_\ell W_{n,\ell}$. The contradiction we're after is to show that $\hat{T}_n$ thus defined is also a good estimator of T_n (given by Equation (A1) with $p = p^\star$), so we need to compare $\min_\ell W_{n,\ell}$ and T_n and show that they are close.

Let us first fix one instance of ℓ. For clarity, let us momentarily drop the ℓ-notation from subscripts. For one fixed such instance, note that W_n is a continuous tail that is coupled with T_n: they are related through the common values of X. For notational convenience, let $X_{\max} := \max_i X_i$ and let $Y_{\max} = \max_i Y_i$. Now observe that T_n is equal to $\sum_{x > X_{\max}} p(x)$ and it is exactly the $F^\star$-measure of the interval $[X_{\max} + 1, \infty)$, by construction. On the other hand W_n is the $F^\star$-measure of the interval $[Y_{\max}, \infty)$. Since $Y_{\max} \in [X_{\max}, X_{\max} + 1]$, it follows that $W_n \geq T_n$. How much larger can it get? By at most $p(X_{\max})$. More formally, we can write $F^\star(x) = \sum_{x'} p(x') G(x - x')$. Thus $W_n - T_n = \int_{Y_{\max}}^{X_{\max}+1} F^\star(dx) = \sum_{x'} p(x') \int_{Y_{\max}-x'}^{X_{\max}+1-x'} G(dx) = p(X_{\max}) \overline{G}(Y_{\max} - X_{\max})$, since the inner integral is non-zero only for $x' = X_{\max}$. Define the (random) set of maximizing observations as $\mathcal{I}_{\max} = \{i : X_i = X_{\max}\}$. Now, let us bound the probability of any particular excess beyond $p(X_{\max})$ times a factor $t \in (0, 1)$:

$$\begin{aligned} \mathbf{P}\{W_n > T_n + p(X_{\max}) \cdot t\} &= \mathbf{P}\left\{ \overline{G}(Y_{\max} - X_{\max}) > t \right\} = \mathbf{P}\left\{ Y_{\max} - X_{\max} < \overline{G}^{-1}(t) \right\} \\ &= \mathbf{P}\left\{ \forall i \in \mathcal{I}_{\max},\ Z_i < \overline{G}^{-1}(t) \right\} \\ &= \mathbf{E}\left[\mathbf{P}\left(\cap_{i \in \mathcal{I}_{\max}} \{Z_i < \overline{G}^{-1}(t)\} \,\middle|\, X_1, \cdots, X_n \right) \right] \\ &= \mathbf{E}\left[\mathbf{P}^{|\mathcal{I}_{\max}|} \{Z < \overline{G}^{-1}(t)\} \right] \\ &\leq \mathbf{P}\{Z \leq \overline{G}^{-1}(t)\} = G(\overline{G}^{-1}(t)) = 1 - t, \end{aligned} \tag{A6}$$

where we conditioned over all X_i, used the independence of the Z_i's from the X_i's to write the probability of the intersection as a product, used the fact that the Z_i have identical distribution to a generic Z, and finally bounded $|\mathcal{I}_{\max}|$ (the only term still depending on the X_i's and thus influence by the outer expectation) by 1. The only loss here is from this replacement. This, however, can be shown to be rather tight. For intuition, note that a geometric-$\frac{1}{2}$ has only a single maximizing observation in expectation, i.e., $\mathbf{E}[|\mathcal{I}_{\max}|] = 1$. This is not good news, since $p(X_{\max})/T_n = \frac{p(X_{\max})}{\sum_{x > X_{\max}} p(x)}$ is lower bounded away from zero in the dithered geometric family, and thus this shows that we cannot expect T_n to be arbitrarily close to a single W_n with probability that is arbitrarily large. This is true *regardless* of the choice of G. This is the motivation behind choosing the smallest of m continuous versions for the

reduction, which restores the needed maneuverability for the approximation. Indeed, now restoring the ℓ-notation:

$$\begin{aligned}
\mathbf{P}\{\min_{\ell=1}^{m} W_{n,\ell} > T_n + p(X_{\max})\cdot t\} &= \mathbf{P}\bigcap_{\ell=1}^{m}\{W_{n,\ell} > T_n + p(x)\cdot t\} \\
&= \mathbf{P}\bigcap_{\ell=1}^{m}\left\{\forall i \in \mathcal{I}_{\max},\ Z_{i,\ell} < \overline{G}^{-1}(t)\right\} \\
&= \mathbf{E}\left[\mathbf{P}\left(\bigcap_{\ell=1}^{m}\bigcap_{i\in\mathcal{I}_{\max}}\{Z_{i,\ell} < \overline{G}^{-1}(t)\}\ \middle|\ X_1,\cdots,X_n\right)\right] \\
&= \mathbf{E}\left[\mathbf{P}^{m|\mathcal{I}_{\max}|}\{Z < \overline{G}^{-1}(t)\}\right] \\
&\leq \mathbf{P}^{m}\{Z \leq \overline{G}^{-1}(t)\} = G(\overline{G}^{-1}(t))^{m} = (1-t)^{m},
\end{aligned} \tag{A7}$$

where the only notable observation is that the m replicated versions compound the number of Zs that deviate. The rest is derived in the same way as in Equation (A6). Finally, using the fact that $\frac{p(X_{\max})}{T_n} \leq \max_x \frac{p(x)}{\sum_{x'>x} p(x')} \leq \kappa$, we have:

$$\mathbf{P}\left\{\frac{\min_{\ell=1}^{m} W_{n,\ell}}{T_n} > 1 + \kappa t\right\} \leq (1-t)^{m}.$$

Specializing to $1+\kappa t = \sqrt{\gamma}$, using the fact that we chose m such that $\left(1 - \frac{\sqrt{\gamma}-1}{\kappa}\right)^{m} < \frac{\gamma-1}{2}$, and recalling that the ratio cannot be less than 1, we have for all n that:

$$\mathbf{P}\left\{\frac{\min_{\ell=1}^{m} W_{n,\ell}}{T_n} \notin \left[\sqrt{\gamma}^{-1}, \sqrt{\gamma}\right]\right\} < \frac{\gamma-1}{2}. \tag{A8}$$

Also specializing Equation (A5) to $\eta = \sqrt{\gamma}$, we have that for n large enough:

$$\mathbf{P}\left\{\frac{\min_{\ell=1}^{m} W_{n,\ell}}{\hat{T}_n} \notin \left[\sqrt{\gamma}^{-1}, \sqrt{\gamma}\right]\right\} \leq \frac{\gamma-1}{2}. \tag{A9}$$

By combining these two approximations, our reduction is complete. Namely, given our assumption (A3) that we can estimate the continuous tail, we have that for n large enough we can estimate the discrete tail to our desired accuracy:

$$\mathbf{P}\left\{\frac{\hat{T}_n}{T_n} \notin [\gamma^{-1}, \gamma]\right\} < \gamma - 1,$$

which clearly contradicts the impossibility (A2) of estimating the discrete tail. More precisely, since the continuous vs. discrete tail approximation in Equation (A8) does *not* depend on the assumption, it must be that it's Equation (A9) that fails for each (n_k). Recalling the bound of Equation (A5), we must have:

$$m\mathbf{P}\left\{\frac{\hat{W}_{n_k}}{W_{n_k}} \notin \left[\sqrt{\gamma}^{-1}, \sqrt{\gamma}\right]\right\} \geq \mathbf{P}\left\{\frac{\min_{\ell} W_{n_k,\ell}}{\hat{T}_{n_k}} \notin \left[\sqrt{\gamma}^{-1}, \sqrt{\gamma}\right]\right\} > \frac{\gamma-1}{2}.$$

Finally, set $\epsilon = \min\left\{\sqrt{\gamma}-1, 1-\sqrt{\gamma}^{-1}, \frac{\gamma-1}{2m}\right\}$ with the above, to obtain the exact claim of the theorem. Namely, for this absolute constant ϵ and subsequence (n_k), given any $\hat{W}_n$ we can construct $F^\star$ as we did (with an arbitrary G), such that $\mathbf{P}_{F^\star}\left\{\frac{\hat{W}_{n_k}}{W_{n_k}} \notin [1-\epsilon, 1+\epsilon]\right\} > \epsilon$ for all k.

Appendix B. This Appendix Presents the Proof of Theorem 4

Appendix B.1. Notation and Outline

Let us first set some notation. Recall that the mean of the geometric distribution $p_\alpha(x) = (1-\alpha)\alpha^{x-1}$ is $\mu = \frac{1}{1-\alpha}$ and its variance is $\sigma^2 = \frac{\alpha}{(1-\alpha)^2}$. Let us write the empirical mean and our parameter estimate respectively as follows:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\alpha}_n = 1 - \frac{1}{\hat{\mu}_n}.$$

The plug-in probability estimate can be expressed as:

$$\check{p}_n(x) := (1-\hat{\alpha}_n)\hat{\alpha}_n^{x-1}.$$

Using our notation for the missing symbols, $E_n := \{x \in \mathbb{N}_+ : \hat{p}(x) = 0\}$, the missing mass is

$$M_n = p_\alpha(E_n) = \sum_{x \in E_n} (1-\alpha)\alpha^{x-1}$$

and the suggested plug-in estimator can be written as

$$\check{M}_n := \check{p}_n(E_n) = \sum_{x \in E_n} (1-\hat{\alpha}_n)\hat{\alpha}_n^{x-1}.$$

The following proof first establishes the convergence of the parameter estimate and then pushes it forward to the entire distribution, specializing in particular to the missing mass. For the latter, we establish some basic localization properties of the punctured segment of a geometric sample coverage. This is related to the general study of gaps, see, for example, [24].

We have the following elementary convergence property for the parameter.

Lemma A1 (Parameter Convergence)**.** *Let $\delta > 0$, and define:*

$$\epsilon_n := \sqrt{\frac{\alpha}{\delta n}} \cdot \left(\frac{\max\{1, \frac{1-\alpha}{\alpha}\}}{1 - \sqrt{\frac{\alpha}{\delta n}}} \right).$$

Then, at every $n > \frac{\alpha}{\delta}$, we have that with probability greater than $1-\delta$:

$$\left| \frac{\hat{\alpha}_n}{\alpha} - 1 \right| \le \epsilon_n \quad \textit{and} \quad \left| \frac{1-\hat{\alpha}_n}{1-\alpha} - 1 \right| \le \epsilon_n.$$

If we let $\eta_n = \epsilon_n/(1-\epsilon_n)$, we can also write this as

$$\frac{1}{1+\eta_n} \le \frac{\hat{\alpha}_n}{\alpha} \le 1+\eta_n \quad \textit{and} \quad \frac{1}{1+\eta_n} \le \frac{1-\hat{\alpha}_n}{1-\alpha} \le 1+\eta_n.$$

Proof. From Chebyshev's inequality, we know that for all $\delta > 0$:

$$\mathbf{P}\left\{ |\hat{\mu}_n - \mu| \le \frac{\sigma}{\sqrt{\delta n}} \right\} \ge 1-\delta.$$

We now simply have to verify that $|\hat{\mu}_n - \mu| \leq \frac{\sigma}{\sqrt{\delta n}}$ implies that both $\left|\frac{\hat{\alpha}_n}{\alpha} - 1\right|$ and $\left|\frac{1-\hat{\alpha}_n}{1-\alpha} - 1\right|$ are smaller than ϵ_n. Indeed, using $\hat{\mu}_n \geq \mu - \frac{\sigma}{\sqrt{\delta n}}$:

$$\left|\frac{\hat{\alpha}_n}{\alpha} - 1\right| = \left|\frac{(\hat{\mu}_n - 1)\mu}{\hat{\mu}_n(\mu - 1)} - 1\right| = \left|(\hat{\mu}_n - \mu)\frac{1}{\hat{\mu}_n(\mu - 1)}\right| \leq |\hat{\mu}_n - \mu| \, \frac{1}{(\mu - \frac{\sigma}{\sqrt{\delta n}})(\mu - 1)}$$

and

$$\left|\frac{1-\hat{\alpha}_n}{1-\alpha} - 1\right| = \left|\frac{\mu}{\hat{\mu}_n} - 1\right| = \left|(\mu - \hat{\mu}_n)\frac{1}{\hat{\mu}_n}\right| \leq |\hat{\mu}_n - \mu| \, \frac{1}{(\mu - \frac{\sigma}{\sqrt{\delta n}})}.$$

Finally, since $|\hat{\mu}_n - \mu| \leq \frac{\sigma}{\sqrt{\delta n}}$, both of these bounds are smaller than:

$$\frac{\sigma}{\sqrt{\delta n}} \frac{1}{(\mu - \frac{\sigma}{\sqrt{\delta n}}) \min\{1, \mu - 1\}} = \frac{\frac{\sqrt{\alpha}}{1-\alpha}}{\sqrt{\delta n}} \frac{1}{(\frac{1}{1-\alpha} - \frac{\sqrt{\alpha}}{1-\alpha}\frac{1}{\sqrt{\delta n}}) \min\{1, \frac{\alpha}{1-\alpha}\}},$$

which is equal to ϵ_n. The expression with η_n follows from $1 - \epsilon_n = \frac{1}{1+\eta_n}$ and $1 + \eta_n > 1 + \epsilon_n$. □

It follows from Lemma A1 that with probability greater than $1 - \delta$, we have the following pointwise convergence of the distribution.

$$(1 + \eta_n)^{-x}(1 - \alpha)\alpha^{x-1} \leq \check{p}_n(x) \leq (1 + \eta_n)^{x}(1 - \alpha)\alpha^{x-1}.$$

Since the rate of this convergence is not uniform, we need to exercise care when specializing to particular events. We focus on the missing symbols' event. We have:

$$\frac{\sum_{x \in E_n}(1 + \eta_n)^{-x}(1 - \alpha)\alpha^{x-1}}{\sum_{x \in E_n}(1 - \alpha)\alpha^{x-1}} \leq \frac{\check{M}_n}{M_n} = \frac{\check{p}_n(E_n)}{p_\alpha(E_n)} \leq \frac{\sum_{x \in E_n}(1 + \eta_n)^{x}(1 - \alpha)\alpha^{x-1}}{\sum_{x \in E_n}(1 - \alpha)\alpha^{x-1}}. \tag{A10}$$

The event E_n is inconvenient to sum over, because it has points spread out randomly. This is particularly true for its initial portion, where the samples "puncture" it. It is more convenient to approximate this segment in order to bound Equation (A10). We now formalize this notion, via the following definition.

Definition A1 (Punctured Segment)**.** *The punctured segment of a sample is the part between the end of the first contiguous coverage and the end of the total coverage. Its extremities are:*

$$V_n^- := \min E_n \quad \textit{and} \quad V_n^+ := \max E_n^{\mathsf{c}}.$$

We have the following localization property for the punctured segment of samples from a geometric distribution.

Lemma A2 (Localization of Punctured Segment)**.** *Let $X_1, \cdots, X_n$ be samples from a geometric distribution $p_\alpha(x) = (1 - \alpha)\alpha^{x-1}$ on $\mathbb{N}_+$. Let V_n^- and V_n^+ be the extremities of the punctured segment as defined in Definition A1. Then, for all $u > (\frac{\alpha}{1-\alpha})^2$, we have:*

$$\mathbf{P}\{V_n^- < \log_{1/\alpha}(n) - \log_{1/\alpha}(u)\} < 2e^{-\frac{1-\alpha}{\alpha}u} < \frac{\alpha}{(1-\alpha)u},$$

$$\mathbf{P}\{V_n^+ > \log_{1/\alpha}(n) + 1 + \log_{1/\alpha}(u)\} < \frac{1}{u}.$$

In particular, for $\delta < (1 - \alpha)/\alpha^2$, we have that with probability greater than $1 - \delta$:

$$\log_{1/\alpha}(n) - \log_{1/\alpha}\left[\tfrac{1}{(1-\alpha)\delta}\right] \leq V_n^- < V_n^+ \leq \log_{1/\alpha}(n) + 1 + \log_{1/\alpha}\left[\tfrac{1}{(1-\alpha)\delta}\right].$$

Proof. Given an integer $a \geq 2$, the event that $V_n^- < a$ implies that at least one of the symbols below a did not appear in the sample. By using the union bound, we thus have that:

$$\begin{aligned}
\mathbf{P}\{V_n^- < a\} &\leq \sum_{x=1}^{a-1} \left[1-(1-\alpha)\alpha^{x-1}\right]^n \\
&= \sum_{\ell=1}^{a-1} \left[1-\frac{(1-\alpha)n\alpha^{a-1-\ell}}{n}\right]^n \\
&\leq \sum_{\ell=1}^{a-1} \exp\left[-(1-\alpha)n\alpha^{a-1-\ell}\right] \leq \sum_{\ell=1}^{\infty} \exp\left[-(1-\alpha)n\alpha^{a-1-\ell}\right]
\end{aligned}$$

By specializing to $a(u,n) = \lfloor \log_{1/\alpha}(n) + 1 - \log_{1/\alpha}(u) \rfloor$:

$$\begin{aligned}
\mathbf{P}\{V_n^- < \log_{1/\alpha}(n) - \log_{1/\alpha}(u)\} &\leq \mathbf{P}\{V_n^- < a(u,n)\} \\
&\leq \sum_{\ell=1}^{\infty} \exp\left[-(1-\alpha)n\alpha^{\log_{1/\alpha}(n)-\log_{1/\alpha}(u)-\ell}\right] \\
&= \sum_{\ell=1}^{\infty} \exp\left[-(1-\alpha)\alpha^{-\ell}u\right].
\end{aligned}$$

Lastly, if $u > (\frac{\alpha}{1-\alpha})^2$, one can show by induction that $(1-\alpha)\alpha^{-\ell}u \geq \frac{1-\alpha}{\alpha}u + \ell - 1$. This turns the sum into a geometric series, giving:

$$\mathbf{P}\{V_n^- < \log_{1/\alpha}(n) - \log_{1/\alpha}(u)\} \leq \mathrm{e}^{-\frac{1-\alpha}{\alpha}u} \sum_{\ell=1}^{\infty} \mathrm{e}^{-\ell+1} < 2\mathrm{e}^{-\frac{1-\alpha}{\alpha}u} < \frac{\alpha}{(1-\alpha)u}.$$

Next, note that V_n^+ is nothing but the maximum of the samples. Thus, given an integer $b \in \mathbb{N}_+$, the event $V_n^+ > b$ is the complement of the event that all the samples are at b or below. Since the total probability of the range $1, \cdots, b$ is $1-\alpha^b$, we thus have:

$$\mathbf{P}\{V_n^+ > b\} = 1 - (1-\alpha^b)^n.$$

If we now specialize to $b(u,n) = \lceil \log_{1/\alpha}(n) + \log_{1/\alpha}(u) \rceil$, we have that:

$$\begin{aligned}
\mathbf{P}\{V_n^+ > \log_{1/\alpha}(n) + 1 + \log_{1/\alpha}(u)\} &\leq \mathbf{P}\{V_n^+ > b(u,n)\} \\
&\leq 1 - \left(1 - \alpha^{\log_{1/\alpha}(n)+\log_{1/\alpha}(u)}\right)^n \\
&= 1 - \left(1 - \frac{1}{u \cdot n}\right)^n < \frac{1}{u}.
\end{aligned}$$

For the last part of the claim, we let $u = \frac{1}{(1-\alpha)\delta}$, followed by a union bound on the analyzed events. This gives us that at least one of the two events holds with probability at most $\frac{1}{u} + \frac{\alpha}{(1-\alpha)u} = \delta$, and therefore neither holds with probability at least $1-\delta$, as desired. □

Appendix B.2. Completing the Proof

We now put together the pieces of the proof of Theorem 4. To show that our estimator learns the missing mass in relative error with respect to $\mathcal{G}$, we obtain the following equivalent statement. Fix $\delta > 0$ and $\eta > 0$. We prove that for n large enough with probability greater than $1-2\delta$ we have:

$$\frac{1}{1+\eta} < \frac{\check{M}_n}{M_n} < 1+\eta.$$

Without loss of generality, to satisfy the conditions of Lemmas A1 and A2, we restrict ourselves to $\delta < (1-\alpha)/\alpha^2$ (we can always choose a smaller δ than specified) and $n > \frac{\alpha}{\delta}$ (we can always ask for n to be larger). As such, we have that with probability at least $1-2\delta$, both events of Lemmas A1 and A2 occur. We work under the intersection of these events.

We give the details of only the right tail of the convergence; all the steps can be directly paralleled for the left tail. To see why the punctured set is a useful notion, we claim that the following quantity upper bounds the right tail of Equation (A10):

$$\begin{aligned}\frac{\sum_{x>V_n^+}(1+\eta_n)^x(1-\alpha)\alpha^{x-1}}{\sum_{x>V_n^+}(1-\alpha)\alpha^{x-1}} &= (1+\eta_n)^{V_n^+}\frac{\sum_{y\in\mathbb{N}_+}(1+\eta_n)^y(1-\alpha)\alpha^{y-1}}{\sum_{y\in\mathbb{N}_+}(1-\alpha)\alpha^{y-1}=1}\\ &= (1+\eta_n)^{V_n^+}\frac{(1-\alpha)(1+\eta_n)}{1-\alpha(1+\eta_n)}. \end{aligned} \tag{A11}$$

where for the first equality we have used the change of variable $y = x - V_n^+$ and simplified the common α factors in the numerator and denominator, and for the second equality we have used the moment generating function of the geometric distribution: $\mathbf{E}[e^{sX}] = (1-\alpha)e^s/(1-\alpha e^s)$. To prove this claim, we proceed by induction, starting at step $t = 1$ with the set $G^{(1)} := \{V_n^+ + 1, V_n^+ + 2, \cdots\} \subset E_n$, adding at every step t the largest element $z^{(t)}$ of E_n not yet in $G^{(t-1)}$ to obtain $G^{(t)}$, and proving that:

$$\frac{\sum_{x\in G^{(t)}}(1+\eta_n)^x(1-\alpha)\alpha^{x-1}}{\sum_{x\in G^{(t)}}(1-\alpha)\alpha^{x-1}} \le \frac{\sum_{x\in G^{(t-1)}}(1+\eta_n)^x(1-\alpha)\alpha^{x-1}}{\sum_{x\in G^{(t-1)}}(1-\alpha)\alpha^{x-1}}.$$

We use the following basic property that for positive real numbers a_1, b_1, a_2, b_2, the following three equalities are equivalent (these are *mediant inequalities*):

$$\begin{aligned}(i) \qquad a_1/b_1 &\le a_2/b_2,\\ (ii) \qquad a_1/b_1 &\le (a_1+a_2)/(b_1+b_2),\\ (iii) \qquad (a_1+a_2)/(b_1+b_2) &\le a_2/b_2.\end{aligned}$$

For the base case, let $a_2 = \sum_{x\in G^{(1)}}(1+\eta_n)^x(1-\alpha)\alpha^{x-1}$ and $b_2 = \sum_{x\in G^{(1)}}(1-\alpha)\alpha^{x-1}$. We then choose the largest $z^{(1)} \in E_n \setminus G^{(1)}$ and we let $a_1 = (1+\eta_n)^{z^{(1)}}(1-\alpha)\alpha^{z^{(1)}-1}$ and $b_1 = (1-\alpha)\alpha^{z^{(1)}-1}$. From (A11), noting that the fraction is always greater than 1, it follows that $a_2/b_2 > (1+\eta_n)^{V_n^+} > (1+\eta_n)^{z^{(1)}} = a_1/b_1$. We can thus add $z^{(1)}$ to the sum, and obtain $(a_1+a_2)/(b_1+b_2) \le a_2/b_2$, establishing the base case. Please note that this also shows that $(a_1+a_2)/(b_1+b_2) \ge a_1/b_1 = (1+\eta_n)^{z^{(1)}}$. We pass this property down by induction, and we can assume this holds true at every step.

To continue the induction at step t, let $a_2 = \sum_{x\in G^{(t-1)}}(1+\eta_n)^x(1-\alpha)\alpha^{x-1}$ and $b_2 = \sum_{x\in G^{(t-1)}}(1-\alpha)\alpha^{x-1}$. As noted, we assume that $a_2/b_2 \ge (1+\eta_n)^{z^{(t-1)}}$ from the previous induction step. We then choose the largest $z^{(t)} \in E_n \setminus G^{(t-1)}$ and we let $a_1 = (1+\eta_n)^{z^{(t)}}(1-\alpha)\alpha^{z^{(t)}-1}$ and $b_1 = (1-\alpha)\alpha^{z^{(t)}-1}$. Since $z^{(t-1)} < z^{(t)}$, it follows that $a_2/b_2 \ge (1+\eta_n)^{z^{(t-1)}} > (1+\eta_n)^{z^{(t)}} = a_1/b_1$. We can thus add $z^{(t)}$ to the sum, and obtain $(a_1+a_2)/(b_1+b_2) \le a_2/b_2$, as desired. Note that this also shows that $(a_1+a_2)/(b_1+b_2) \ge a_1/b_1 = (1+\eta_n)^{z^{(t)}}$, and the induction is complete.

By combining this result with the equivalent argument on the left side, we have shown that we can replace Equation (A10) by

$$\frac{\sum_{x\ge V_n^-}(1+\eta_n)^{-x}(1-\alpha)\alpha^{x-1}}{\sum_{x\ge V_n^-}(1-\alpha)\alpha^{x-1}} \le \frac{\check{M}_n}{M_n} = \frac{\check{p}_n(E_n)}{p_\alpha(E_n)} \le \frac{\sum_{x>V_n^+}(1+\eta_n)^x(1-\alpha)\alpha^{x-1}}{\sum_{x>V_n^+}(1-\alpha)\alpha^{x-1}}$$

or equivalently by

$$(1+\eta_n)^{-V_n^-+1}\frac{(1-\alpha)(1+\eta_n)^{-1}}{1-\alpha(1+\eta_n)^{-1}} \le \frac{\check{M}_n}{M_n} \le (1+\eta_n)^{V_n^+}\frac{(1-\alpha)(1+\eta_n)}{1-\alpha(1+\eta_n)}. \tag{A12}$$

In Lemma A1 we have set:

$$\eta_n = \epsilon_n/(1-\epsilon_n),$$

with

$$\epsilon_n := \sqrt{\frac{\alpha}{\delta n}} \cdot \left(\frac{\max\{1, \frac{1-\alpha}{\alpha}\}}{1 - \sqrt{\frac{\alpha}{\delta n}}} \right).$$

On the other hand, by Lemma A2, we have that:

$$V_n^+ \leq \log_{1/\alpha}(n) + 1 + \log_{1/\alpha}\left[\frac{1}{(1-\alpha)\delta}\right]$$

and

$$V_n^- \geq \log_{1/\alpha}(n) - \log_{1/\alpha}\left[\frac{1}{(1-\alpha)\delta}\right].$$

It follows that both bounds of Equation (A12) converge to 1, at the rate of roughly $\log(n)/\sqrt{n}$, instead of the parametric rate $1/\sqrt{n}$. Regardless, for any desired $\eta > 0$, we get that there exists a large enough n beyond which, with probability greater than $1 - 2\delta$, we satisfy:

$$\frac{1}{1+\eta} \leq \frac{\check{M}_n}{M_n} \leq 1 + \eta.$$

This establishes that $\check{M}_n$ learns M_n, as desired.

References

1. Good, I.J. The population frequencies of species and the estimation of population parameters. *Biometrika* **1953**, *40*, 237–264. [CrossRef]
2. McAllester, D.A.; Schapire, R.E. *On the Convergence Rate of Good–Turing Estimators*; COLT: Hartford, CT, USA, 2000; pp. 1–6.
3. McAllester, D.A.; Ortiz, L.E. Concentration inequalities for the missing mass and for histogram rule error. *J. Mach. Learn. Res.* **2003**, *4*, 895–911.
4. Berend, D.; Kontorovich, A. On the concentration of the missing mass. *Electron. Commun. Probab.* **2013**, *18*, 1–7. [CrossRef]
5. Ohannessian, M.I.; Dahleh, M.A. Rare Probability Estimation Under Regularly Varying Heavy Tails. *JMLR Proc.* **2012**, *23*, 21.1–21.24.
6. Ben-Hamou, A.; Boucheron, S.; Ohannessian, M.I. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli* **2017**, *23*, 249–287. [CrossRef]
7. Taleb, N.N. *The Black Swan: The Impact of the Highly Improbable*; Random House: London, UK, 2008.
8. Antos, A.; Lugosi, G. Strong minimax lower bounds for learning. *Mach. Learn.* **1998**, *30*, 31–56. [CrossRef]
9. Beirlant, J.; Goegebeur, Y.; Segers, J.; Teugels, J. *Statistics of Extremes: Theory and Applications*; Wiley: Hoboken, NJ, USA, 2004.
10. Beirlant, J.; Devroye, L. On the impossibility of estimating densities in the extreme tail. *Stat. Probab. Lett.* **1999**, *43*, 57–64. [CrossRef]
11. Rajaraman, N.; Thangaraj, A.; Suresh, A.T. Minimax risk for missing mass estimation. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 3025–3029.
12. Acharya, J.; Bao, Y.; Kang, Y.; Sun, Z. Improved Bounds for Minimax Risk of Estimating Missing Mass. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 326–330.
13. Orlitsky, A.; Suresh, A.T. Competitive distribution estimation: Why is Good–Turing good. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2018; pp. 2143–2151.
14. Valiant, G.; Valiant, P. Instance optimal learning of discrete distributions. In Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing, Cambridge, MA, USA, 19–21 June 2016; pp. 142–155.
15. Valiant, G.; Valiant, P. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new CLTs. In Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing, San Jose, CA, USA, 6–8 June 2011; pp. 685–694.

16. Wu, Y.; Yang, P. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *arXiv* **2015**, arXiv:1504.01227.
17. Orlitsky, A.; Suresh, A.T.; Wu, Y. Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 13283–13288. [CrossRef]
18. Falahatgar, M.; Ohannessian, M.I.; Orlitsky, A.; Pichapati, V. The power of absolute discounting: All-dimensional distribution estimation. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; pp. 6660–6669.
19. Boucheron, S.; Gassiat, E.; Ohannessian, M.I. About adaptive coding on countable alphabets: Max-stable envelope classes. *IEEE Trans. Inf. Theory* **2015**, *61*, 4948–4967. [CrossRef]
20. Kneser, R.; Ney, H. Improved smoothing for m-gram language modeling. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Detroit, MI, USA, 9–12 May 1995; pp. 679–682.
21. MacKay, D.; Peto, L. A hierarchical Dirichlet language model. *Nat. Lang. Eng.* **1995**, *1*, 289–307. [CrossRef]
22. Pitman, J.; Yor, M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **1997**, *25*, 855–900. [CrossRef]
23. Teh, Y.W. A hierarchical Bayesian language model based on Pitman-Yor processes. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL), Sydney, Australia, 17–18 July 2006; pp. 985–992.
24. Louchard, G.; Prodinger, H. On gaps and unoccupied urns in sequences of geometrically distributed random variables. *Discret. Math.* **2008**, *308*, 1538–1562. [CrossRef]